A Multi-View Camera-Projector System for Object Detection and Robot-Human Feedback

Jinglin Shen, Jingfu Jin and Nicholas Gans

Abstract—In this paper, we present a novel camera-projector system for assisting robot-human interaction. The system is comprised of a stereo camera pair and a DLP projector. The proposed system provides feedback information indicating the robot's perception of the environment and what action a human user desires. Feedback is delivered by iteratively spotlighting objects in the environment using the projector. When the spotlight lands on an object that the human user wants the robot to retrieve, he or she can confirm the object selection, and the robot will perform a grasping task to retrieve the selected object. In this investigation, the proposed camera-projector performs three tasks: 1) Actively detect visually salient objects in the scene from the two camera views using a visual attention model. 2) Locate the detected objects using a trifocal tensor based object matching method, and project a spot of light on the objects to provide information back to the human user. 3) Reconstruct the 3D position of the target to plan robot motion and conduct vision-based control to move a robot manipulator to grasp the object. Experimental results of human-directed object grasping task are presented to demonstrate the functions of the proposed system.

I. INTRODUCTION

In the near future, robots will be increasingly used for human service and assistance applications in the home. Examples of recent applications include service robots that can assist people eating and shaving [1], a mobile robotic follower to carry oxygen tanks [2], and unloading items from a dishwasher [3]. Many of these applications involve object approaching, grasping and manipulation, which often requires either a controlled environment with known 3D structures or an efficient way to acquire information about the target object in order to plan robot motion.

To extend the abilities of service robots, human-robot interaction and human-in-the-loop systems draw increasing research attention. Such systems allow human operators to convey their commands or intentions to robots using natural interfaces like gesture [4] and voice [5]. In other cases, human input is used to help robot better understand the environment, for example in [6], the operator uses voice commands to enhance segmentation of occluding objects in the image views.

While natural interfaces for human-robot interaction receives extensive study, it appears that less attention has been paid to how robots convey their views of the environment or tasks to the human operator in a natural way. One approach would be to display information that the robots obtain about object locations, shapes, sizes etc. directly on or near the objects rather than on a computer monitor. The major goal of this work is to develop a camera-projector system that enables robots to provide such instructional feedback. Our motivating example is an object grasping task. The system projects a spotlight on different objects that are detected, which it believes a human user might want it to retrieve. When the spotlight arrives on a desired object, the human user can confirm selection and the robot grasps the object.

We present a camera-projector system and demonstrate a human-robot interaction experiment based on it. To locate objects that would likely interest a human user, we modify Itti’s visual attention model [7] by adding depth information as a channel to the saliency map and adapt a weighting scheme based on information theory that was proposed in [8]. Then, the detected objects are mapped from camera views to the projector view to provide feedback information. This requires first matching the objects in the camera views. A novel object matching method based on the trifocal constraints is proposed. Lighting from the projector is used, but in a way that differs from traditional structured light methods. In the literature, various algorithms have been developed for stereo matching and 3D reconstruction based on stereo depth maps [9], [10]. Structured light methods generally work better for objects that lack visual texture or features [11]. In recent years, the Kinect device, which uses infrared structured light to achieve depth measurement, has gained a lot of research interests in the area of human-computer interaction [12], [13]. Recently, Li, et al., proposed to combine the stereo matching and structured light methods to compensate for their drawbacks [14]. However, most existing stereo matching methods are either not accurate or too slow to use in real time.

The system works as follow. Object detection is first done
in an bottom-up manner by using the visual attention model. Objects are then segmented in the saliency maps of the two camera views separately. Segmented objects in the two camera views are matched by projecting a spotlight, whose position is calculated from trifocal constraints for each pair of matching candidates. Correct matching will cause the object to be illuminated by the spotlight, which can be detected by the cameras. Shape and area comparison is done ahead of the matching process to reduce the number of possible matches for each object. By including depth information in the saliency map, objects that appears occluded in the camera views but at different depth can be separated. After matching each segmented object in camera views, the projector projects a spotlight on all matched objects one by one and waits for human input. Once a human command is received, further grasping operation can be performed. The structure of the system is illustrated in Fig. 1.

A closely related work is done by Rasolzadeh et al. [15]. They developed a vision attention based active vision system with four cameras to guild robot grasping tasks. However, the object segmentation and 3D reconstruction is based on dense stereo matching. Another major difference is that our system is capable of providing feedback information through natural interface by directly projecting on the objects.

The paper is organized as follows. Related background is given in Section II. The system calibration procedure is introduced in Section III. Section IV describes how the systems functions are realized, including objection detection based on the visual attention model and trifocal constraints, illuminating the detected objects and stereo reconstruction for visual servoing. Then we demonstrate the experimental result in Section V, and finally Section VI gives the conclusion and future works.

II. BACKGROUND

A. Camera and Projector Model

The coordinates of a 3D point are defined as \( M = [X, Y, Z, 1]^T \) using homogeneous coordinates measured in the inertial world reference frame. Its corresponding 2D projection is defined as \( m = [x, y, 1]^T \) as measured in the camera frame. Under the pinhole camera model, the projection relation between \( M \) and \( m \) is given by

\[
m = PM = \begin{bmatrix}
\frac{1}{Z} & 0 & 0 & 0 \\
0 & \frac{1}{Z} & 0 & 0 \\
0 & 0 & 1 & 0
\end{bmatrix}
\begin{bmatrix}
R & t \\
0 & 1
\end{bmatrix} M,
\]

where \( R \in S O(3) \) and \( t \in \mathbb{R}^3 \) are rotation and translation of the camera frame with respect to the world origin. \( P \in \mathbb{R}^{3 \times 4} \) is defined as the projection or camera matrix, and \( Z \) is the \( Z \) coordinate of the point as measured in the camera frame. The homogeneous coordinates \( m \) is converted into coordinates in image plane by

\[
p = Km,
\]

where \( K \in \mathbb{R}^{3 \times 3} \) is the calibration matrix of the camera.

A projector can be modeled as a dual of a pinhole camera, with the difference being the direction of projection. That is, for a camera, a point on a 3D surface in space projects to a 2D image point. While for a projector, a 2D image point \( m \) is projected as a line \( L \) in 3D space (i.e. \( L \) is the preimage of \( m \)). The intersection of the line \( L \) with a surface results in a point of light on the surface. To distinguish between the directions of projection, we define projecting from 3D space to 2D image plane to be “projecting in” (i.e. projection of camera), and projecting from 2D image to 3D space to be “projecting out” (i.e. projection of projector).

B. Trifocal Tensor

The trifocal tensor encodes the geometric relation among three camera views and is analogous to the role of the fundamental matrix in two-view geometry. Since the pinhole model of a projector is reciprocal to a camera, the geometry of the trifocal tensor relation is still valid if any of three cameras is replaced by a projector. Therefore, we will talk about general three-view systems in the following discussion.

The trifocal tensor \( T \) can be presented using a set of three matrices as \( T = \{T_i, T_2, T_3\} \), where \( T_i \in \mathbb{R}^{3 \times 3} \) for \( i \in \{1, 2, 3\} \). According to [16], given three fixed projection matrices \( P = [I, 0] \), \( P' = [A, a] \) and \( P'' = [B, b] \) (without loss of generality, the world frame is assumed to align with one camera frame, as defined by projection matrix \( P \)), the trifocal tensor is uniquely defined by

\[
T_i = a_i b_4^T - a_i b_i^T,
\]

where \( a_i \) and \( b_i \) are the \( i_{th} \) column of \( P' \) and \( P'' \), \( i \in \{1, 2, 3\} \). Conversely, if the trifocal tensor is known, projection matrices \( P' \) and \( P'' \) can be recovered as

\[
P' = [T_1 e'' T_2 e'' T_3 e' T_e'] \quad (3)
\]

\[
P'' = [(e'' e'^T) T_1^T e', (e'' e'^T) T_2^T e', (e'' e'^T) T_3^T e', e''], \quad (4)
\]

where \( e' \) and \( e'' \) are the epipoles in the views defined by \( P' \) and \( P'' \). The epipoles \( e' \) and \( e'' \) can also be solved from the trifocal tensor as the left null-vectors of matrices \( [u_1, u_2, u_3] \) and \( [v_1, v_2, v_3] \) where \( u_i \) and \( v_i \) are the left and right null-vectors of \( T_i \) respectively. Note that (2) to (4) correspond to the case when \( P = [I, 0] \) (i.e. the world origin is fixed at the center of \( P \)). Detailed analysis and the trifocal tensor in the general case is given in [16].

Like the epipolar constraint in epipolar geometry, there are constraints among corresponding points and lines in the three views for the trifocal tensor, namely incidence relations. There are several incidence relations, including line-line-line correspondence and point-line-line correspondence, etc. In this work, the point-point-point correspondence relation is used for solving the trifocal tensor, and the point-line-point correspondence is used to locate matches of the selected points. The point-point-point correspondence is illustrated in Fig. 2(a), and the relation defines a constraint expressed using Matrix notation as

\[
[m]_x \sum_i T_i m'' = 0_{3 \times 3}, \quad i \in \{1, 2, 3\}
\]

where \( [m]_x \) refers to the \( 3 \times 3 \) skew-symmetric matrix made from vector \( m \).

The point-line-point correspondence is illustrated in Fig. 2(b). The preimage \( \pi' \) of line \( l' \) in view 2 defines a plane, which induces a homography between the image points \( m \).
and \( m'' \) in the other two views. Therefore, given a point \( m \) in view 1 and a line \( l' \) in view 2 which passes through \( m' \), the image point \( m'' \) can be found in view three by

\[
m'' = H_{13}(l')m = (\sum_i m_i T_i)'', \quad i \in \{1, 2, 3\}
\]

(6)

where \( H_{13} \) is the homography matrix induced by \( l' \), and \( m_i \) is the \( i_{th} \) term of \( m \). Note that \( l' \) can be any line that passes through \( m' \) except the epipolar line, which corresponds to a degenerate case of the trifocal tensor.

### III. System Calibration Using the Trifocal Tensor

In this section, the calibration procedure of the camera-projector system is discussed. Given fixed relative positions among the three views, the calibration procedure results in projection matrices \( P, P' \) and \( P'' \). Calibration of such a camera-projector system involves intrinsic calibration of both cameras and the projector, and extrinsic calibration among the three views using the trifocal tensor. Intrinsic calibrations are performed first. Camera intrinsic calibration is done using the Matlab Camera Calibration Toolbox [17]. Projector intrinsic calibration is done using the method of Falcao et al. [18], which is also based on [17].

After intrinsic calibration of the cameras and projector, we apply the Gold Standard algorithm [16] to calculate the trifocal tensor and recover the projection matrices. Image correspondences are acquired by using the projector to display a checkerboard pattern of known size on several different planar surfaces. The inner corners of the checkerboard squares are extracted from the camera views to get \( n > 7 \) sets of feature points. Using the captured images correspondences, an initial linear solution can be calculated according to the point-point-point relation in (5). A nonlinear optimization step then follows using the Levenberg-Marquardt method, taking the linear solution as a starting point.

### IV. Camera-Projector System Functions

The major functions of the system include finding and matching objects using a saliency map and trifocal constraints, providing feedback information to the human operator and recovering 3D structure of the scene.

#### A. Object Detection based on the Visual Attention Model

In the saliency-based visual attention model, attractiveness of visual stimuli is represented as a saliency map, which models visual response to a given scene. Several methods have been proposed for object detection based on visual attention [7], [19]. Regions of an image with high saliency will likely be of interest to a human viewing the scene.

In addition to the common saliency channels such as color, intensity and orientation, depth information is considered an important factor that affects human visual attention in some applications. The idea of adding depth information as a channel first appeared in [20]. Depth information and stereo disparity is used in various ways in [21], [22].

In this work, we add a disparity map as a feature map to Itti’s saliency model and weight the feature maps using an information theory based scheme. In recent years, the Microsoft Kinect has been a popular device for generating depth maps using structured light. While it is an feasible option, the projector in the system is used to project structured light patterns so that no extra device is needed. Using the Kinect to replace the stereo camera in the system is one direction of our future work.

Generating a disparity map is a well studied problem in computer vision. In this work, we adapt a structured light method using temporally encoded binary stripe patterns [23]. Directly adding the disparity map as a channel to the saliency map will promote some background regions after normalization, since there are regions in which depth information cannot be recovered or have false matches, as shown in Fig. 3(b). Therefore some pre-processing is done on the disparity map to remove the background. In an indoor environment with a background wall like the scene in Fig. 3 (a), the disparity of the background wall corresponds to the maximum value in the histogram of the disparity map in Fig. 3 (b). Removing the background from the disparity map results in Fig. 3 (c). In more general cases with no clear background, the disparity constraints determined by the maximum range of the projector and the robot arm can be used to define a threshold, and any disparity value under this threshold should be suppressed.

When combining the feature maps into the final saliency map, the feature maps with fewer regions standing out from the background should be favored. We adapt the weighting
scheme proposed in [8]. The probability that an event is observed from a feature map $M$ is defined as

$$p(M) = \sum_{i,j} M_{\tau}(i,j),$$

where $M_{\tau}(i,j)$ is the pixel value higher than a conspicuous threshold $\tau$, and $M(i,j)$ is a pixel in the map $M$. The weight of the feature map is decided by the amount of information obtained from the map

$$I(M) = -\log(p(M)).$$

The final saliency map is given by the combination of the four feature maps in this manner:

$$S = \sum_{i=1}^{4} I(M_i)M_i.$$

The saliency maps generated by this weighting scheme are shown in Fig. 3(d)-(f) when using no disparity map, the disparity map directly, and the disparity map after background removal respectively. It is can be seen from Fig. 3(d) that without the disparity map included, it is difficult to separate any object on the table from the background. By adding depth information in Fig. 3(e), the green tea box can be distinguished from the jar in the back which appears occluded by the tea box in the camera view. However, the white/red cup on the left remains undistinguished from the pile of books behind it, and the objects on the wall become less salient. Saliency map with a disparity map after the background removal in Fig. 3(f) shows that all the objects can be easily segmented from the background.

B. Trifocal Tensor Based Object Matching

In order to project feedback information on detected objects, they need to be located in the projector view. This requires that the objects be matched in the two camera views. We propose an object matching method based on the trifocal tensor constraints that makes use of lighting from the projector. Regions with high saliency values are first segmented from the saliency maps. The centroids of the contours that represent the salient areas are extracted as candidates for matching.

The matching procedure works as follows. Let the sets of all centroids in the two camera views be $C$ and $C'$ respectively, and define subsets $\tilde{C} = \{c, c \in C, c \neq e\}$ and $\tilde{C}' = \{c', c' \in C', c' \neq e'\}$, where $e$ and $e'$ are the epipoles. If $e = e'$ or $e' = e''$, the corresponding point in the projector view can not be found. This is because the preimage of $e$ and $e'$ are collinear, so the intersection is undefined, and the 3D point can not be located. For any point $c \in \tilde{C}$ and $c' \in \tilde{C}'$, a point $c''$ can be found in the projector view according to the point-line-point incidence in (6) by constructing a line $l'$ that passes $c'$ but not $e'$. Let $M$ be a point in space; its images points $m$, $m'$ and $m''$ in the three views are uniquely defined by the trifocal constraints. If $c = m$, the point $c''$ found by the point-line-point correspondence coincides with $m''$ only when $c' = m'$. Projecting a pattern out at $e''$, the preimage passes $M$, which will cause the pattern to be detected at location $c$ and $c'$ in the camera views. Otherwise, if $c' = m'' \neq m'$, the point-line-point correspondence maps $e''$ to a different point $m''$, whose preimage does not pass $M$. In this work, we iterate through each pair of $c \in C$ and $c' \in C'$, project out a white dot, and check the intensity change at point $c$ in the left camera view. When intensity change is observed, a match between $c$ and $c'$ is found for an object. This is illustrated in Fig.4.

Some pre-processing can be done to reduce the number of possible matches for an object before the matching procedure. In this work, shape and area of the segmented regions are used, assuming that the two cameras are located close to each other and their facing angles are similar. For each point $c \in C$, a subset $S' \subseteq C'$ can be defined as all $c' \in C'$ that satisfies the following criterion:

$$|area(r) - area(r')| < T_a,$$

$$S(r, r') < T_s,$$

where $T_a$ and $T_s$ are tunable thresholds, $r$ and $r'$ are the segmented regions whose centroids are $c$ and $c'$, and $S$ is the similarity score defined by

$$S(r, r') = \sum_{i=1}^{7} \frac{1}{|log(h'_i)|} - \frac{1}{|log(h''_i)|},$$

where $h'_i$ and $h''_i$ are the $i$th Hu moments of $r$ and $r'$. The criterion means only regions with similar areas and shapes are considered for matching. Then $c$ is matched with $c' \in S', \frac{S(r, r')}{S(r, r')} > T_a$, which usually reduces the number of possible matches to 1 or 2 in our experiments. A complete example of object matching is shown in Fig. 5.

C. Presenting Feedback Information Using the Projector

Once the camera-projector system detects one or more salient objects in the environment, it can provide information back to the human by projecting certain patterns of light on the objects. The feedback information can be used to indicate the system’s knowledge of the positions, shapes, sizes of the detected objects and so on, or to convey its planned task for human confirmation when an object is selected through natural interfaces. The process of projecting patterns on to detected objects is similar to the trifocal tensor based method for object matching described in Section IV-B.

For each object matched in the two camera views, the bounding ellipses are found, and the corresponding points of
the ellipse center, vertices and co-vertices in the projector view are calculated using the point-line-point correspondences. Therefore, the corresponding ellipse in the projector view can be drawn using these feature points. Projecting the pattern out will highlight an elliptical region on the target object that covers much of the 3D object. This ensures that a human user will see the spotlight. It is possible to project various meaningful patterns, such as arrows, numbers, colors etc. In this work, the pattern of fitting ellipse to the object contour is used to maximize the area being spotlighted on the object, and the benefits of different shapes and colors will be investigated in future works.

By using the trifocal tensor based matching method, 3D feedback information can be determined using only the 2D relations of the three image views, without 3D reconstruction of the entire scene. Note that if depth information is used as a channel in the saliency map, stereo matching may be already available depending on the method used to generate the disparity map. However, most accurate stereo matching methods are very slow, and real time stereo matching methods are often not accurate or dense enough. Therefore, the depth information is sufficient in building the saliency map, but may be not reliable for mapping the feedback pattern to the projector view accurately. On the other hand, the proposed object matching relies on only the 2D mapping between the three views, and guarantees the correct matching regardless of the accuracy of disparity reconstruction.

D. 3D Reconstruction and Object Grasping

When a grasping task is specified, 3D reconstruction can be done locally within the region of the target object. The size of the target object and its 3D coordinates in the world frame are obtained from stereo reconstruction using the feature points of the bounding ellipses. The gripper attached to the robot arm is detected in the two camera views using a fiducial marker, and the 3D coordinates of the gripper are recovered. The goal position of the robot gripper is set to a grasping position behind the object. Closed-loop visual servoing proceeds [24], with the robot end effector arriving stably at the goal position to grasp the object. Only position error is considered in this work, and advanced grasping planning [25], including approach direction and angle will be addressed in future works.

V. EXPERIMENTAL RESULTS

In this section, we present experimental results of using the proposed camera-projector system to detect everyday objects in a typical indoor environment (i.e. objects have no fiducial markers and the background is somewhat cluttered). Based on human selection, the system conducts a visual servoing control task to guide the robot to grasp the target object and retrieve it for the user.

A. Calibration Accuracy

Accurate calibration is critical for the system functions to work correctly. Simulation of the Gold Standard method for general three-view systems is given in [16]. The camera-projector system consists of a DLP and a Minoru Stereo webcam. The projector is mounted on a tripod, and the stereo camera is rigidly attached on top of the projector. Note this is for convenience, it is not necessary that the camera system be located so close to the DLP.

A total of 20 feature points are obtained by projecting a checkerboard pattern onto a planar surface at different positions. The calibration accuracy is evaluated by the reprojection error in the three views. In a typical experiment, the reprojection error of the initial linear solution is 0.278 pixels, and the error after applying the Levenberg-Marquardt algorithm is 0.068 pixels. The result shows that the gold standard method recovers geometric relations of the three views with very small reprojection error, which indicates the calibration is accurate.

B. Object Detection and Projecting Feedback Information

The experiment is performed in a normal lab environment with lightly cluttered background. Several objects are placed on a table in front of the camera-projector system. The objects are at different distances to the system, and the wall is approximately 4 meters away. The range of the projector implicitly defines a region of interest. The objects outside of the projector range cannot be illuminated, and therefore cannot be matched using the proposed matching method.
The experiment proceeds as follows. The system first projects a sequence of structured light patterns to build the disparity map. Saliency maps are then generated for the incoming images from the two cameras, and salient objects are segmented in the two views. The system then iterates through all the objects in one view and looks for the matching object in the other view using the criterion described in Section IV-B. The estimated object centroids $c$, and $c'$ generally do not coincide exactly with the actual matching images $m, m'$ of the a point $M$ due to parallax and the image error produced by estimating centroids $c$ and $c'$. To ensure the preimage by projecting $c''$ out intersects with $M$ on the object, a solid circle of diameter 30 pixels is projected instead of a single point at $c''$.

Once an object is identified, the system projects an elliptical feedback pattern on it for 5 seconds while waiting for human feedback. After all the objects are checked, the system repeat the object detection and matching process. Because the process is completely automated, the system can adapt to dynamic changes of the scene and movement of the system.

Some experimental results on a sample scene are shown in Fig. 6. Four objects are detected and matched in the given scene, including the black eraser, the stuffed toy, the tea box and the fire alarm on the wall. The robot arm is also detected using the visual attention model, but fails to match in the two camera views, since the estimated $c''$ is outside of the projector image range and the arm cannot be lightened by the projected circle.

One potential problem of projecting elliptical patterns is that novice users can confuse it with the light projection for object matching. The size of the elliptical pattern help differentiate, and we have experimented with different light colors. More meaningful feedback patterns as suggested in section IV-C can be used to distinguish from the object detection light. Another solution is to use devices like Microsoft Kinect, which utilize infrared projector and camera for object matching, and only represent feedback patterns using the DLP projector.

C. Object Grasping Guided by the Camera-Projector System

We present the experimental result of object grasping task with assistance of the proposed camera-projector system. A six degree of freedom Staubli TX90 robot arm is used. Position based visual servoing is performed to move the robot arm to the grasping position. The visual servoing solely relies on the 3D reconstruction and position error calculation achieved by the camera-projector system. A fiducial marker is attached to the left side of the gripper, and is located 60 mm from the center of the gripper. The position of the marker is tracked using the Lucas-Kanade method.

In this experiment, only position error is considered for visual servoing, and a fixed goal position is used. More advanced grasping planning will be studied in future works. The goal position is set to be 60 mm to the left of the object center, half the height of the object below the object top, and 20 mm behind the target. Position error is measured in the left camera frame. Relative pose between the camera and end effector at its home position is calibrated off-line using a calibration board.

Object detection and feedback proceed. A human operator pressed the button after the black eraser was highlighted, and the robot arm moved to the desired goal position to grasp the eraser. The feedback gains of the PBVS controller are tuned such that the error along the vertical direction converges slower than the other two directions. This ensures that the end effector moves above and behind the object quickly and approaches the object from the top but does not bump it. The initial position and final position after convergence is shown in Fig. 7. The position error over time as measured in the left camera frame, is given in Fig. 8, which shows that the visual servoing converges based on 3D reconstruction using the camera-projector system. A complete object grasping experiment can be seen in the included video.

In summary, the experiments show that after calibration using the proposed procedure, the camera-projector system is capable of automatically finding salient objects in a random lab environment, and providing accurate feedback information using trifocal constraints to direct a spotlight. Robot applications such as object grasping can be done based on the 3D reconstruction using the system.
VI. CONCLUSION AND FUTURE WORKS

We present a camera-projector system for robot-human interaction. A calibration procedure based on the trifocal tensor is used to accurately calibrate the three-view system. An automated object detection method using a bottom-up visual attention model with disparity map is proposed, and objects in the two camera views are matched using the trifocal tensor constraints. Projecting feedback information and 3D reconstruction of the target object is achieved to realize robot-human interaction through a natural interface. An object grasping experiment guided by the system is done, which shows that the system is able to actively detect visually salient objects, convey its perception to the human user, and assist grasping tasks to retrieve the selected target.

There are several open avenues for future works. We are replacing the stereo camera with a Kinect RGBD camera to provide depth information, and simplify the geometric relation for mapping objects into the projector view. With depth information, the current visual attention model can distinguish objects close to the system from those that appear occluded but located far away. In more complex scenes that objects are close or next to each other, advanced segmentation like the graph cut based methods are needed to better separate them. Grasping planning including path planning and grasping direction will be studied as a direct application of the proposed system.

REFERENCES


Fig. 7. Initial and grasping position of the robot arm. The gripper fiducial and target center are highlighted with green dots.

Fig. 8. Position error converges along all three axes.